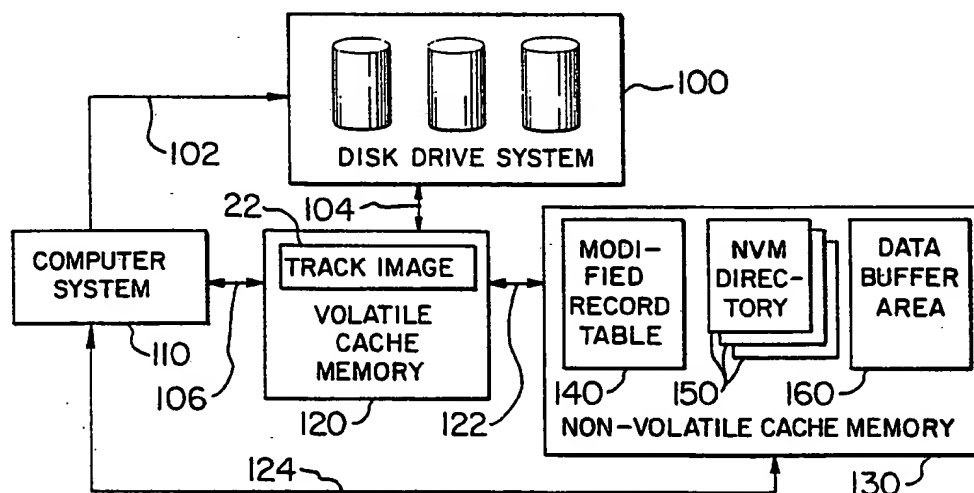




INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification ⁵ : G06F 12/08	A1	(11) International Publication Number: WO 93/21579 (43) International Publication Date: 28 October 1993 (28.10.93)
(21) International Application Number: PCT/US92/03299 (22) International Filing Date: 21 April 1992 (21.04.92) (71) Applicant: STORAGE TECHNOLOGY CORPORATION [US/US]; 2270 South 88th Street, Louisville, CO 80028 (US). (72) Inventors: KURZAWA, Leonard, Joseph ; 10315 Owens Circle, Broomfield, CO 80021 (US). PETERSON, Gregory, William ; 2567 South Star Route, Lyons, CO 80540 (US). (74) Agent: DUFT, Donald, M.; Dorr, Carson, Sloan & Peterson, 3010 East 6th Avenue, Denver, CO 80206 (US).		(81) Designated States: AU, CA, JP, European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IT, LU, MC, NL, SE). Published With international search report.

(54) Title: METHOD FOR MANAGING DATA RECORDS IN A CACHED DATA SUBSYSTEM WITH NON-VOLATILE MEMORY



(57) Abstract

A method is described for managing data records stored in non-volatile memory in a disk drive system with cache memory. A variable-length directory containing descriptors of disk records is used to locate a selected record non-volatile memory. A table, ordered sequentially by record number, is used to quickly locate a record in non-volatile memory without having to perform a time-consuming search. In order to efficiently utilize space in non-volatile memory, a list is kept of free space for storing record descriptors. After an initial nominal allocation, additional free space is allocated only when required, thus further increasing the efficiency of use of non-volatile memory.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	FR	France	MR	Mauritania
AU	Australia	GA	Gabon	MW	Malawi
BB	Barbados	GB	United Kingdom	NL	Netherlands
BE	Belgium	GN	Guinea	NO	Norway
BF	Burkina Faso	GR	Greece	NZ	New Zealand
BG	Bulgaria	HU	Hungary	PL	Poland
BJ	Benin	IE	Ireland	PT	Portugal
BR	Brazil	IT	Italy	RO	Romania
CA	Canada	JP	Japan	RU	Russian Federation
CF	Central African Republic	KP	Democratic People's Republic of Korea	SD	Sudan
CG	Congo	KR	Republic of Korea	SE	Sweden
CH	Switzerland	KZ	Kazakhstan	SK	Slovak Republic
CI	Côte d'Ivoire	LJ	Liechtenstein	SN	Senegal
CM	Cameroon	LK	Sri Lanka	SU	Soviet Union
CS	Czechoslovakia	LJ	Luxembourg	TD	Chad
CZ	Czech Republic	MC	Monaco	TG	Togo
DE	Germany	MG	Madagascar	UA	Ukraine
DK	Denmark	ML	Mali	US	United States of America
ES	Spain	MN	Mongolia	VN	Viet Nam
FI	Finland				

**METHOD FOR MANAGING DATA RECORDS IN A
CACHED DATA SUBSYSTEM WITH NON-VOLATILE MEMORY**

FIELD OF THE INVENTION

This invention relates generally to computer disk
5 drive systems with associated cache memory and non-
volatile memory subsystems, and more particularly, to
a method for efficiently managing write operations to
disk, in a disk drive system in which updated track
images are kept both in volatile cache memory and in
10 non-volatile cache memory.

PROBLEM

Many disk drive systems have associated cache
memory to increase the speed with which data stored in
the disk system can be accessed. Generally, cache
15 memory is volatile; that is, any data stored in cache
is lost if a power failure occurs. Data which has
been modified or updated by a computer system is
vulnerable to such a loss of power between the time
the data is written to cache memory and the time it is
20 written back to the disk drive system. In order to
ensure the integrity of this updated data that has
been written to cache memory, a copy of each disk
record that has been updated is also written to non-
volatile cache memory (NVM), so that it will remain
25 intact in the event of a power failure.

Each track image staged to cache memory is fully

-2-

represented in cache memory with all records on the track being stored in cache memory sequentially, by record number. However, the only records which must be copied into NVM are those records which have been

5 modified (as the result of a write operation) after the track was staged from disk into cache memory. Once a track has been staged in cache memory, records contained in the track may be randomly written to, or modified by, a computer system. This means that NVM

10 will generally contain a variable number of out-of-sequence records. This situation makes NVM difficult to manage in a manner which is efficient and which requires a minimum amount of system time and overhead.

-3-

SOLUTION

The present invention employs a novel method for managing records stored in non-volatile memory. As in the prior art, a directory containing record descriptors (pointers to NVM records) is used to locate a selected record in NVM. However, the method of the present invention uses a variable-length directory to save space in NVM. This reduces the cost of the disk drive system because non-volatile memory is relatively more costly than volatile types of memory. In addition, the prior art required searching through the directory of randomly-ordered descriptors in order to locate a given record in NVM. In contrast, the present invention uses a table of ordered directory entries to quickly locate a desired record, without having to perform a time-consuming search. The table is ordered sequentially by record number, and each entry therein contains the position, in the directory, of the record descriptor corresponding to a given record number. Thus, for example, record number n in NVM can be quickly located in two steps. First, the nth entry in the table is directly addressed. This entry indicates the location of the record descriptor for record number n, which descriptor is then directly addressed. The descriptor contains the address of record n, which has now been found without conducting a sequential search of the record descriptor directory.

Furthermore, since records in NVM are generally re-written back to disk in record number order, the sequentially ordered table provides for an effective method of performing this re-write operation.

In order to efficiently utilize the NVM space, a "free list" of available ("free") space for storing

-4-

record descriptors is also kept in NVM. After an initial nominal allocation, additional free space is allocated only when required, thus further efficiently using non-volatile memory.

BRIEF DESCRIPTION OF THE DRAWING

Figure 1 shows, in block diagram form, the interrelationship between the disk drive system, an associated volatile cache memory subsystem, a track
5 image read from the disk drive system into the volatile cache memory subsystem, the non-volatile cache memory subsystem, and a computer system;

Figure 2 shows the non-volatile memory subsystem containing a plurality of non-volatile memory
10 directories, each containing a plurality of segments for storing record descriptors; and a data buffer area containing a plurality of records that have been modified by the computer system;

Figure 3 shows the non-volatile memory subsystem containing a modified record table partitioned into a
15 plurality of record descriptor lists, and containing a corresponding plurality of non-volatile memory directories;

Figure 4 is a flowchart showing the steps
20 associated with the storing of a copy of a record in non-volatile memory; and

Figure 5 shows a record descriptor list in the modified record table, an associated non-volatile memory directory, and the data buffer area.

-6-

DETAILED DESCRIPTION

Figure 1 shows, in block diagram form, the interrelationship between the disk drive system 100, an associated cache memory subsystem 120, a track image 22 read from the disk drive system 100 into the volatile cache memory subsystem 120, a non-volatile cache memory subsystem 130, and a computer system 110. Note that the non-volatile memory subsystem 130 contains a modified record table 140, a plurality of non-volatile directories 150, and a data buffer area 160, the contents of which are described below.

Whenever the computer system 110 requests that data be written to or read from the disk drive system 100, the computer system 110 transmits information over line 102 identifying the location of the desired data within the disk drive system 100, usually specifying drive number (if the disk drive system 100 has more than one drive), cylinder, track, and record number. Upon issuance of a request by the computer system 110 to read a record from the disk drive system 100, an image 22 of the track containing the requested record is staged (read) from the disk drive system 100 over bus 104 into the volatile cache memory subsystem 120. The computer system 110 can then access the retrieved record in the cache memory subsystem 120 over bus 106.

After a track image has been read into cache memory 120, as records are updated, or modified, by computer system 110, space in a non-volatile memory data buffer 160 is allocated for a copy of these updated records. The computer system 110 stores an updated copy of the requested record in the cache memory subsystem 120 over bus 122, and also stores, over bus 124, for the purpose of data integrity, a

-7-

copy of the modified record or changes thereto in the non-volatile memory data buffer 160. Therefore, the record changes are safeguarded in the event of a power failure between the time the record has been modified by the computer system 110 and the time the record is written from the cache memory subsystem 120 back to the disk drive system 100. A segment 210 of non-volatile memory 130 is also allocated for managing and organizing the record with respect to other records located on the same track image 22.

When a write operation requested by the computer system 110 is complete, with the modified record having been successfully written from the cache memory subsystem 120 back to the disk drive system 100, all space used in non-volatile memory 130 for storage and management of the track associated with the record is deallocated.

If the write operation was unsuccessful because of a data integrity problem, the track image 22 is restaged into volatile cache memory 120 and a copy of the modified record in non-volatile memory 130 is written to the cache memory subsystem 120, and the write to disk operation is attempted again.

Non-Volatile Memory Architecture

Figure 2 shows the non-volatile memory subsystem 130 containing a plurality of non-volatile memory directories 150, each containing a plurality of segments 210, 210' for storing record descriptors 212; and a data buffer area 160 containing a plurality of updated records 113 read from the computer system 110. After a track image 22 has been read from the disk drive system 100 into the cache memory subsystem 120, when a record on the track is updated by the computer

-8-

system 110, a segment 210 of non-volatile memory 130 is initially allocated as an area for storing a plurality of record descriptors 212, each of which contains information pertaining to records in the track image 22. More specifically, each record descriptor 212 references a modified disk record 113 that has been written to the data buffer area 160 in non-volatile memory 130. A record descriptor 212 includes: (1) a flag(s) field indicating whether the associated record has been modified, (2) a pointer to a copy of the record in non-volatile memory, and (3) a "count key data field", containing the cylinder and track number of the record in the disk drive system 100, a record identification number, and the length of the record.

The non-volatile memory segment 210 contains a control area 214 and space for 24 record descriptors 212, with subsequent segments 210' being allocated as necessary, each segment having sufficient space to store 32 record descriptors, because the subsequent segments do not require additional control areas. A control area 214 contains data used to manage an area of memory, and is a concept well-known in the art. Every allocated segment 210 or 210' is the same size. Typically, the 24 record descriptors 212 in the initial segment 210 are sufficient for a given track, but an exceptional track may require up to five segments 210, 210' to be allocated. It should be noted that more or less space could be allocated for a segment 210, 210', depending on the specific characteristics of the disk drive system 100 and the program being run on the computer system 110. Only the required number of segments 210, 210' are allocated for a given track.

-9-

All segments 210 and 210' for a given track are located in a variable size directory 150, in non-volatile memory 130, containing pointers to records which have been modified. This modified data record
5 directory is referred to as a non-volatile memory directory 150. When the computer system 110 requests that a record be updated, a non-volatile memory directory 150 is allocated for that track.

Modified Record Table

10 Figure 3 shows the non-volatile memory subsystem 130 also containing a modified record table 140 partitioned into a plurality of record descriptor lists 350, and containing a corresponding plurality of
15 non-volatile memory directories 150. After a non-volatile memory directory 150 is allocated, an area in non-volatile memory 130, called a modified record table 140, is also allocated for storing indicia of location of record descriptors 212 that are later
20 written to the non-volatile memory directory 150. Each time a track is read from the disk drive system 100, a record descriptor list 350 for that track is created in the modified record table 140. Each record descriptor list 350 is of sufficient size to
25 accommodate 128 entries 354, which number of entries is typically the maximum possible number of records contained on one track of an individual disk in the disk drive system 100 (only several of the 128 entries are shown for purposes of clarity).

All free memory locations (those locations which
30 are not occupied by record descriptors 212) in each non-volatile memory directory 150 are put into a free list 216 in the non-volatile memory directory 150 for the corresponding track. The free list 216 is used

-10-

for storing new record descriptors 354 for the track. Each time a segment 210 or 210' is allocated, a counter, indicating the number of available record descriptors for the corresponding track, is incremented accordingly.

The free list 216 is a singly-linked list with the first available memory location 117 in the free list 216 being pointed to by a head pointer 356 located in the record descriptor list control area 352.

Data Record Writing to Non-Volatile Memory

Figure 4 is a flowchart showing the steps associated with the storing of a copy 113' of record n in non-volatile memory 130 and Figure 5 illustrates the position of data in non-volatile memory 130. When a disk record n on track m is updated as a result of a write operation, the following steps then occur:

(1) At step 402, a copy 113' of record n is written to a data buffer area 160 in non-volatile memory 130;

(2) the free list 216 for track m is checked at step 404 to determine whether there are any available record descriptors 212 in the non-volatile memory directory 150 for track m;

(3) if, at step 406, there is no space remaining in the free list 216 (i.e., there are no available record descriptors 212) for track m, then at step 408 another segment 210' of non-volatile memory is allocated to the free list 216, and the available record descriptor counter (not shown) is incremented by 32 at step 410;

(4) at step 412, a record descriptor 212' for record n is written to the first available location

-11-

504 in the free list 216 in the non-volatile memory directory 150;

5 (5) at step 414, the integer corresponding to the positional offset (rank) Λ of the record descriptor 212' from the beginning of the non-volatile memory directory 150 is then written to the nth location 502 in the record descriptor list 350' for track m in the modified record table 140;

10 (6) the available record descriptor counter in the free list 216 is decremented, at step 416; and

(7) at step 418, the free list head pointer 352 is adjusted to point to the next entry in the free list 216.

15 Note that step (5) above accomplishes the ordering of the records for a given track in non-volatile memory 130.

Record Location in Non-Volatile Memory

20 Figure 5 shows a record descriptor list 350' in the modified record table 140, an associated non-volatile memory directory 150, and the data buffer area 160. In order to locate a modified record 113', in non-volatile memory 130, with a record number n, on track number m, the record descriptor list 350' for track m is first located in the modified record table 140. The nth entry 502 in the record descriptor list 350' is then located. This nth entry 502 contains the offset Λ , into the non-volatile memory directory 150, of the entry 504 containing the record descriptor 212' for the record to be located. The offset Λ is the
30 number of entries from the beginning of the non-volatile memory directory 150 to the location of the desired modified record descriptor 212'. The entries in the record descriptor list 350' and in the non-

-12-

volatile memory directory 150 are directly located (directly addressed) using memory address arithmetic methods well known in the art. The modified record 113' is found by referring to the entry 504 in the non-volatile memory directory 150, which entry 504 contains the address of the record 113' in non-volatile memory 130.

It is to be expressly understood that the claimed invention is not to be limited to the description of the above disclosed preferred embodiment but encompasses other modifications and alterations within the scope and spirit of the inventive concept.

-13-

WE CLAIM:

1. In a disk drive system which stores data in a plurality of tracks, each of said tracks containing a plurality of data records, each of said data records having an associated record number, said disk drive
5 system having an associated volatile memory subsystem and a non-volatile memory subsystem, said non-volatile memory subsystem having random-access memory for storing a plurality of said data records and a plurality of record descriptors, each of said record
10 descriptors comprising a pointer to one of said data records in said non-volatile memory subsystem, a method for managing said data records in said non-volatile memory subsystem, comprising the steps of:
allocating a data buffer area in said non-
15 volatile memory subsystem for storing said data records;
writing, to said data buffer area, a copy of each said data record which has been modified since having been read into said volatile memory subsystem;
20 compiling, in said non-volatile memory subsystem, a directory of said modified data records, each entry in said directory having indicia of location of a corresponding modified data record in said non-volatile memory subsystem;
25 compiling, in said non-volatile memory subsystem, a list for each of said tracks which has been read into said volatile memory subsystem, each of said lists having entries, ordered sequentially by said record number, of the records in said non-
30 volatile memory subsystem, each of said entries in said list having indicia of location of a corresponding said record descriptor in said modified

-14-

data record directory; and

35 locating said selected data record in the
non-volatile memory subsystem, by indexing to the
record descriptor in said list in said non-volatile
memory subsystem, using the record number as an offset
from the beginning of the list, to find an entry in
the list, and then using the value of the found entry
40 as an offset into said modified data record directory
to find the entry having indicia of location of the
selected data record in said non-volatile memory
subsystem.

2. The method of claim 1 wherein:

the step of compiling said lists is
performed in said volatile memory subsystem.

3. The method of claim 1 wherein:

the step of compiling said directory of said
data records includes:

5 allocating a segment of said non-volatile
memory equivalent to the size occupied by the sum of
a pre-determined plurality of said record descriptors.

4. The method of claim 1, further including the
steps of:

5 tabulating, in said non-volatile memory
subsystem, a list of free memory locations available
for storing said record descriptors; and

allocating space for said record descriptors
from the list of free memory locations.

5. The method of claim 5 wherein space for the
record descriptors is allocated in segments capable of
containing between 16 and 64 descriptors per segment.

-15-

6. The method of claim 5 wherein the step of allocating is in response to a request for writing a data record to said disk drive system.

7. The method of claim 6 wherein said segments are allocated only when additional space is required for storing additional record descriptors.

8. In a disk drive system which stores data in plurality of tracks each containing a plurality of data records, each of said data records having an associated record number, said disk drive system
5 having an associated non-volatile memory subsystem having random-access memory for storing a plurality of said data records, said non-volatile memory subsystem containing a plurality of record descriptors, each said record descriptor comprising a pointer to one of
10 said data records in said non-volatile memory subsystem, a method for managing said data records in said non-volatile memory subsystem comprising the steps of:

allocating space for a modified record table
15 in said non-volatile memory subsystem having n entries, each entry being of sufficient size to contain indicia of location of a memory location in said non-volatile memory subsystem, where n is equal to the number of said data records per said track;

20 allocating a segment of space for an non-volatile memory directory, capable of containing v said record descriptors, where v is a variable integral number between one and the number of said data records per said track, said segment being
25 allocated from said free list;

maintaining a free list, in said non-

-16-

volatile memory subsystem, of the unused memory locations in the directory which are available for storing said record descriptors;

30 defining an area in said non-volatile memory subsystem as an non-volatile memory data buffer for storing said data records which are modified;

 writing said data records to the non-volatile memory data buffer;

35 compiling, in the modified record table, for each said track, a list of said record descriptors for said records written to said non-volatile memory subsystem, said record descriptor list having entries sequentially ordered by record number, each said entry
40 in said list having indicia of location of a record descriptor in said non-volatile memory directory, said record descriptor corresponding to one of said data records in said non-volatile memory data buffer; and

 locating said selected data record in said
45 non-volatile memory subsystem, using the record number of said selected data record as an offset from the beginning of said record descriptor list to an entry in said list, then using the integer value of said entry in said list as the rank of position of an entry
50 in said non-volatile memory directory, said non-volatile memory directory entry having indicia of location of said selected record in said non-volatile memory subsystem.

9. The method of claim 8, wherein the step of compiling includes, for each of said data records which has been modified as a result of a request for writing said data record to said disk drive system:

5 storing, in the first available location in the free list, a record descriptor corresponding to

-17-

said modified data record; and

10 storing, in the Nth location in the record descriptor list in the modified record table for the associated track, the integer corresponding to the offset, into the non-volatile memory directory, of the entry containing the record descriptor for said modified data record, where N is the record number of said modified record.

5 10. The method of claim 8 wherein the step of allocating a segment of space for the non-volatile memory directory occurs in response to a request for writing a data record to said disk drive system, when said request requires more space than currently allocated in the free list for the track associated with the record; and v is an integer between 8 and 64.

11. The method of claim 8 wherein the step of allocating space for the modified record table occurs in said volatile memory subsystem.

FIG. 1.

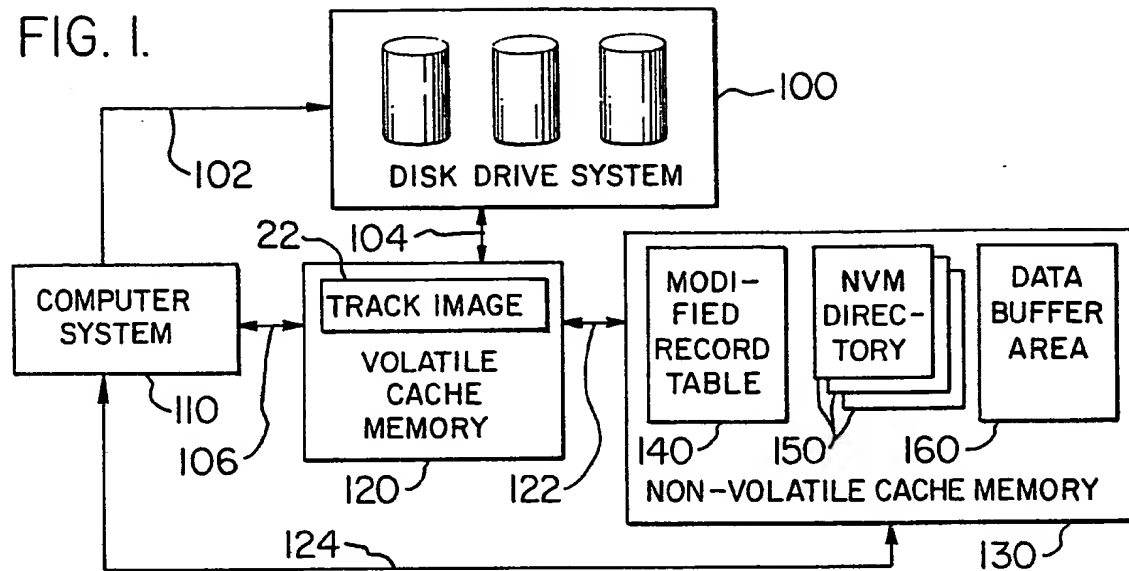
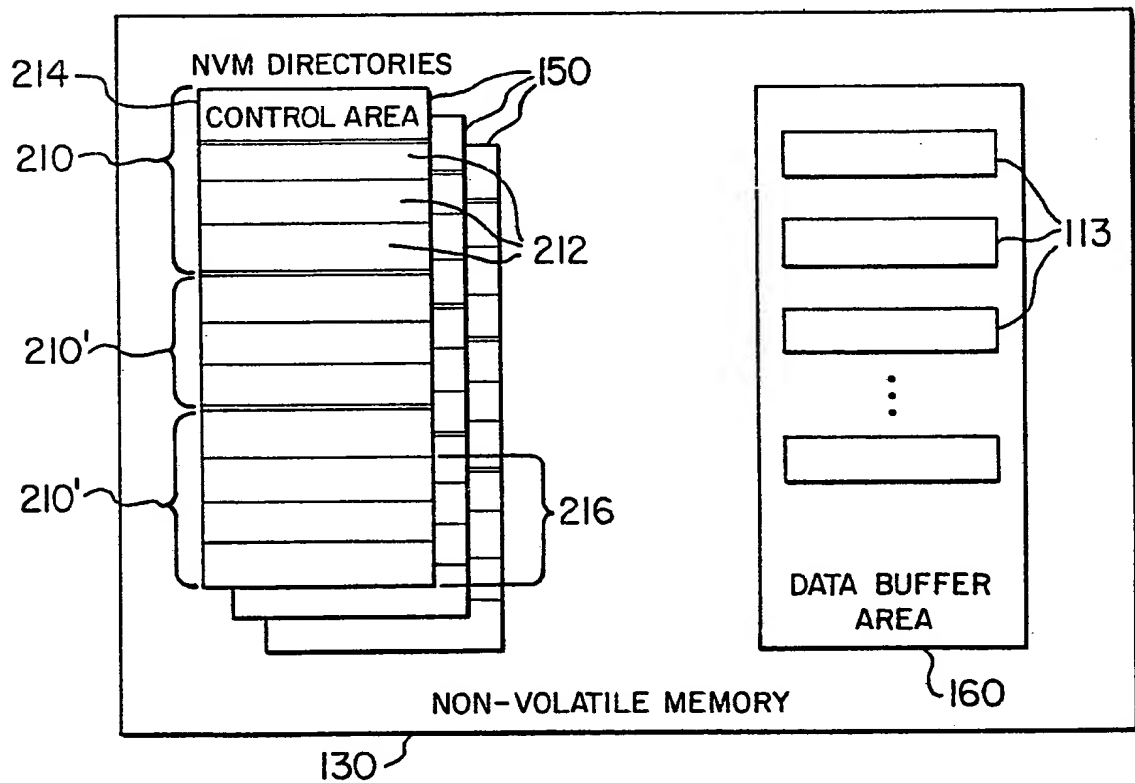
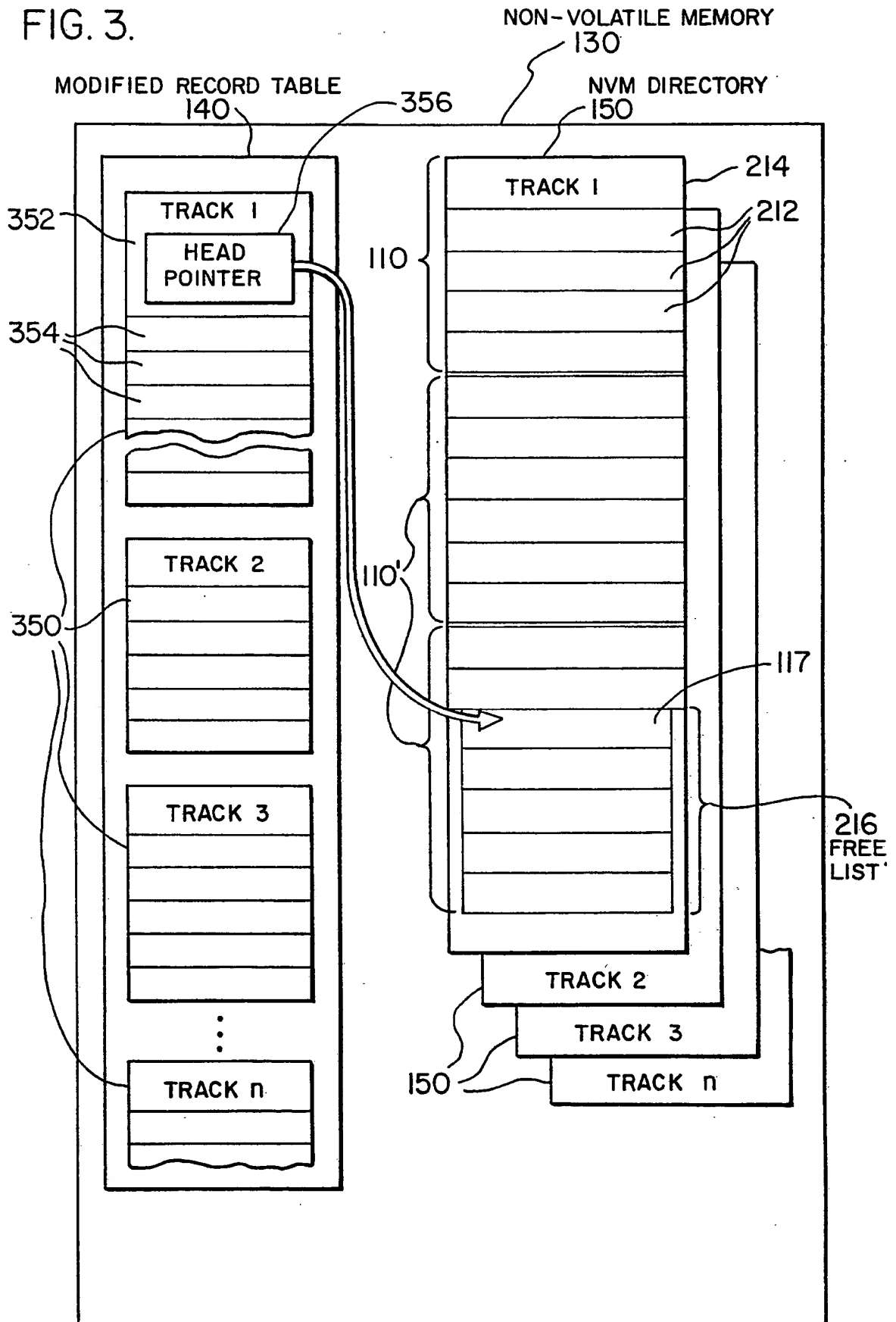


FIG. 2.



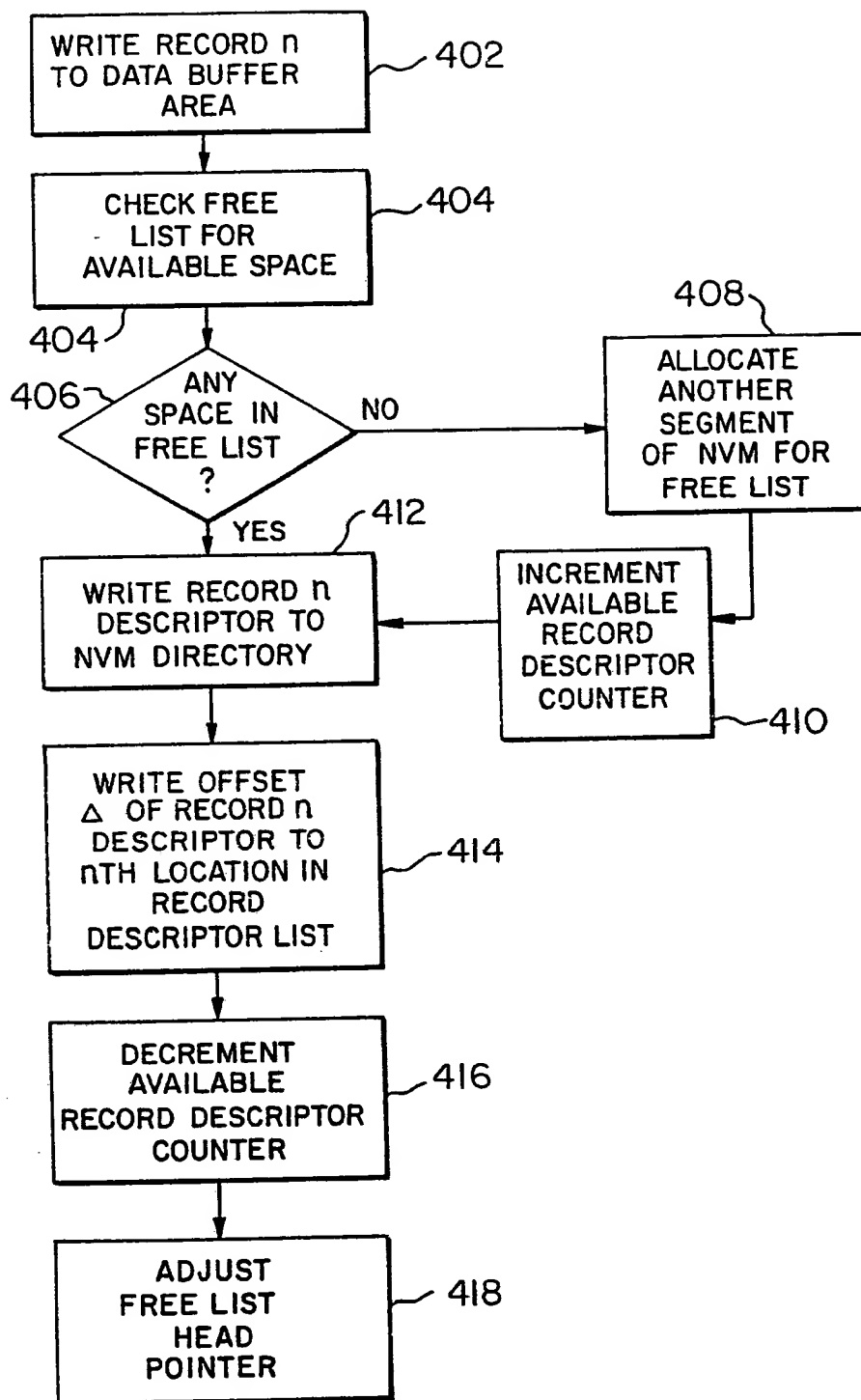
2 / 4

FIG. 3.



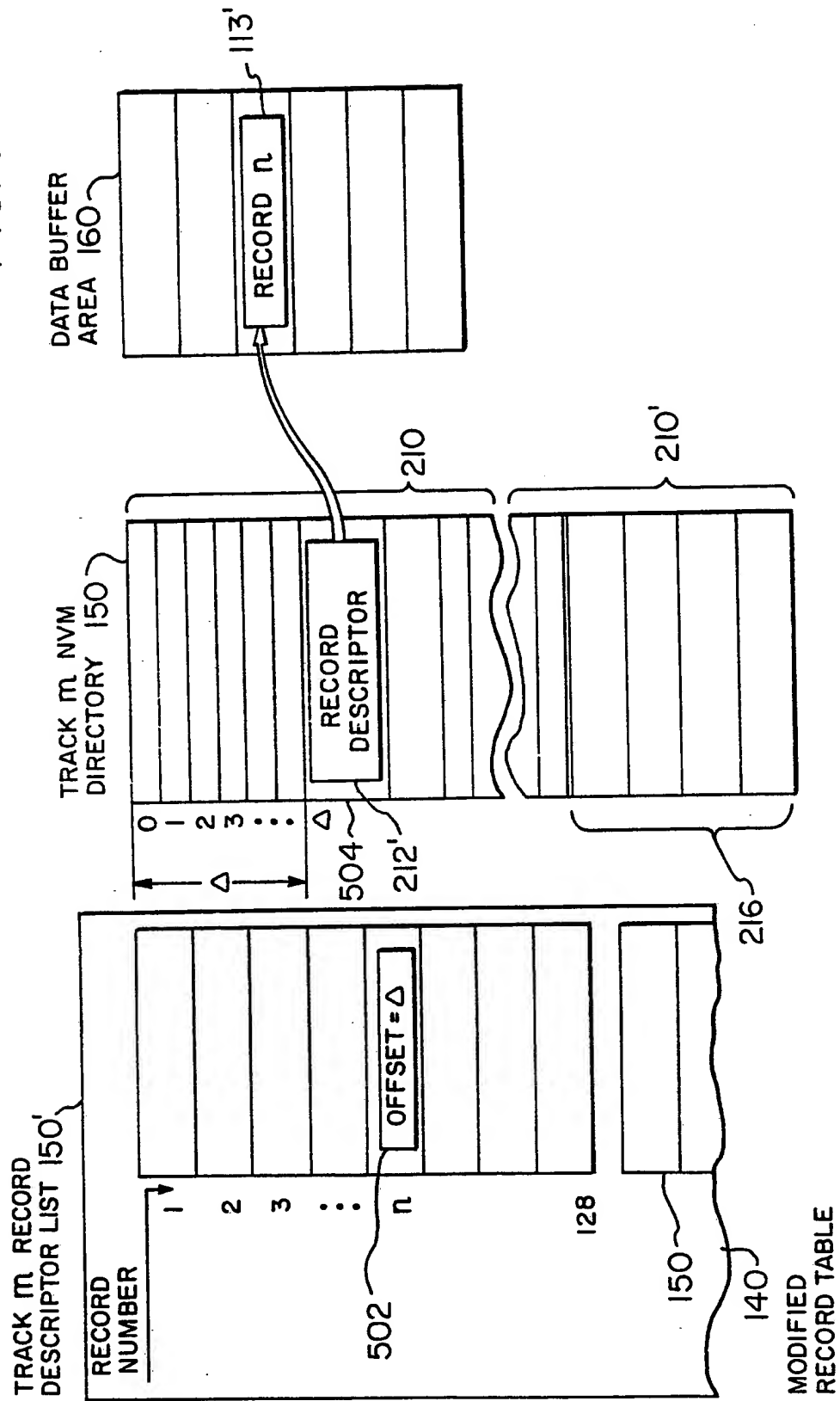
3/4

FIG. 4.



4 / 4


FIG. 5.



INTERNATIONAL SEARCH REPORT

PCT/US 92/03299

International Application No

I. CLASSIFICATION OF SUBJECT MATTER (if several classification symbols apply, indicate all) ⁶		
According to International Patent Classification (IPC) or to both National Classification and IPC Int.Cl. 5 G06F12/08		
II. FIELDS SEARCHED		
Minimum Documentation Searched ⁷		
Classification System	Classification Symbols	
Int.Cl. 5	G06F	
Documentation Searched other than Minimum Documentation to the Extent that such Documents are Included in the Fields Searched ⁸		
III. DOCUMENTS CONSIDERED TO BE RELEVANT⁹		
Category ¹⁰	Citation of Document, ¹¹ with indication, where appropriate, of the relevant passages ¹²	Relevant to Claim No. ¹³
A	IBM TECHNICAL DISCLOSURE BULLETIN. vol. 32, no. 11, April 1990, NEW YORK US pages 81 - 82 , XP97616 'Preventing overflow in a wraparound buffer' see the whole document ---	1,8
A	IBM TECHNICAL DISCLOSURE BULLETIN. vol. 34, no. 2, July 1991, NEW YORK US pages 26 - 31 , XP210554 'Record caching scatter index table directory structure' see the whole document ---	1,8
A	IBM TECHNICAL DISCLOSURE BULLETIN. vol. 33, no. 2, July 1990, NEW YORK US pages 301 - 302 , XP123628 'No modified data indicator for caching subsystem' see the whole document ---	1,8
-/--		
<p>¹⁰ Special categories of cited documents :</p> <p>"A" document defining the general state of the art which is not considered to be of particular relevance</p> <p>"E" earlier document but published on or after the international filing date</p> <p>"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)</p> <p>"O" document referring to an oral disclosure, use, exhibition or other means</p> <p>"P" document published prior to the international filing date but later than the priority date claimed</p> <p>"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention</p> <p>"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step</p> <p>"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.</p> <p>"A" document member of the same patent family</p>		
IV. CERTIFICATION		
Date of the Actual Completion of the International Search	Date of Mailing of this International Search Report	
22 DECEMBER 1992	07.01.93	
International Searching Authority	Signature of Authorized Officer	
EUROPEAN PATENT OFFICE	PFITZINGER E.E. 	

III. DOCUMENTS CONSIDERED TO BE RELEVANT (CONTINUED FROM THE SECOND SHEET)		
Category *	Citation of Document, with indication, where appropriate, of the relevant passages	Relevant to Claim No.
A	<p>PC MAGAZINE vol. 7, no. 17, 11 October 1988, NEW YORK US pages 255 - 260 , XP4538 D. BOLING 'Speed up hard disks with DCACHE' see figure 1</p>	1,8
A	<p>US,A,5 091 909 (KISHIRO ET AL.) 25 February 1992 see figure 3</p>	1,8
A	<p>FUJITSU-SCIENTIFIC AND TECHNICAL JOURNAL vol. 26, no. 4, February 1991, KAWASAKI JP pages 271 - 279 , XP231120 TAKISAWA ET AL. 'F1700 File controller unit' see page 272, right column, last paragraph - page 273, left column see page 275, left column, last paragraph - right column, paragraph I; figure 2</p>	1,8
A	<p>DIGEST OF PAPERS OF THE COMPCON SPRING 1988, FEBRUARY 29 - MARCH 4, SAN FRANCISCO, USA, IEEE NEW YORK, USA pages 146 - 151 , XP212180 J. MENON ET AL. 'The IBM 3990 Disk Cache' see page 148, left column, last paragraph - page 149, left column, paragraph 2</p>	1,8

US 9203299
SA 62354

The European Patent Office is in no way liable for these particulars which are merely given for the purpose of information. 22/12/92

EPO FORM P0479

For more details about this annex : see Official Journal of the European Patent Office, No. 12/82